Box-Cox Transformations: An Overview

Pengfei Li

Department of Statistics, University of Connecticut

Apr 11, 2005

Introduction

Since the seminal paper by Box and Cox(1964), the Box-Cox type of power transformations have generated a great deal of interests, both in theoretical work and in practical applications. In this presentation, I intend to go over the following topics:

- What are the Box-Cox power transformations?
- The inference on the transformations parameter.
- Some cautionary notes on using the Box-Cox transformations.

What are the Box-Cox power transformations?

► The original form of the Box-Cox transformation, as appeared in their 1964 paper, takes the following form:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

► In the same paper, they also proposed an extended form which could accommodate negative y's:

$$y(\boldsymbol{\lambda}) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\ \log(y + \lambda_2), & \text{if } \lambda_1 = 0. \end{cases}$$

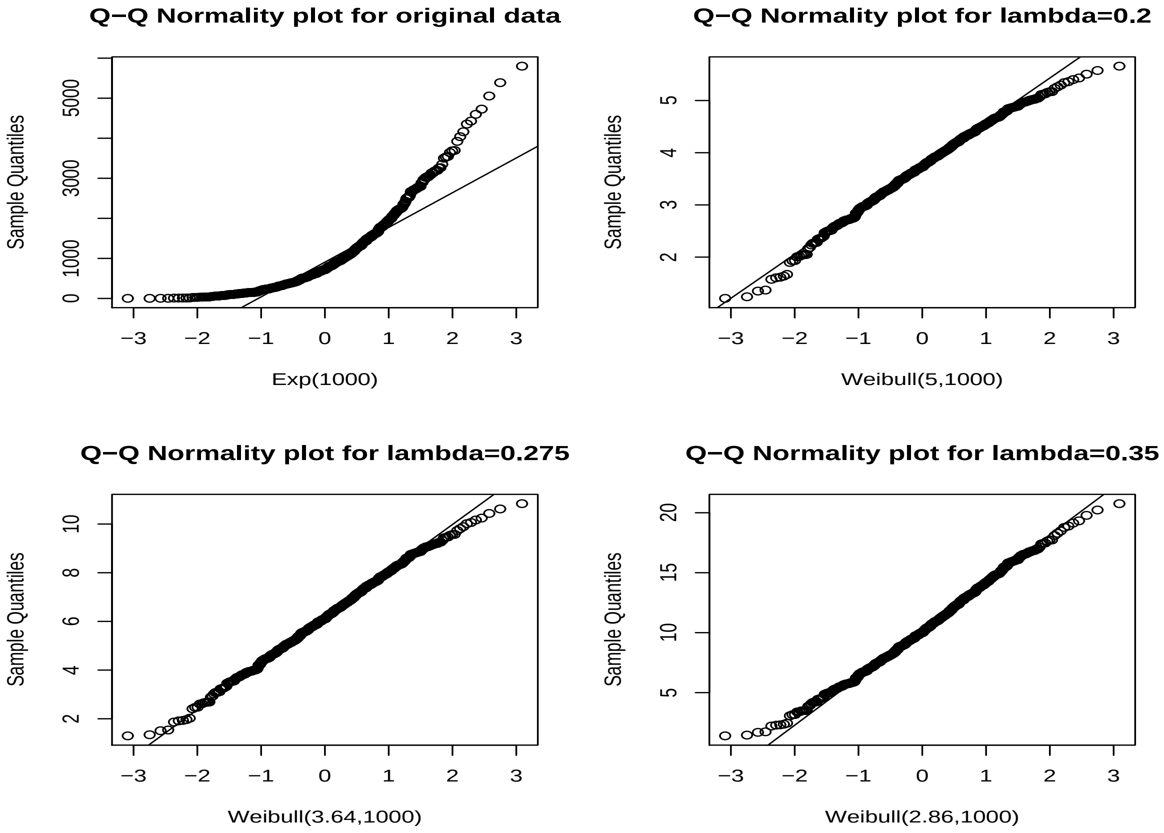
Here, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)'$. In practice, we could choose λ_2 such that $y + \lambda_2 > 0$ for any y. So, we could only view λ_1 as the model parameter.

▶ The aim of the Box-Cox transformations is to ensure the usual assumptions for Linear Model hold. That is, $\boldsymbol{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$

▶ Clearly not all data could be power-transformed to Normal. Draper and Cox (1969) studied this problem and conclude that even in cases that no power-transformation could bring the distribution to exactly normal, the usual estimates of λ will lead to a distribution that satisfies certain restrictions on the first 4 moments, thus will be usually symmetric.

▶ One example in Draper and Cox(1969) is the following: Suppose that the raw data are from an Exp(1000) distribution. The estimate of λ is 0.268. 3 values that are close to 0.268 are chosen to perform the transformation: $\lambda_1 = 0.2, \quad \lambda_2 = 0.275, \quad \lambda_3 = 0.35$. Such transformations result in 3 Weibull distributions: Weib(5,1000), Weib(3.64,1000) and Weib(2.86,1000).

► The following are Q-Q Normal plots for a random sample of size 500 from Exp(1000) distribution.



Since the work of Box and Cox(1964), there have been many modifications proposed.

▶ Manly(1971) proposed the following exponential transformation:

$$y(\lambda) = \begin{cases} \frac{e^{\lambda y}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ y, & \text{if } \lambda = 0. \end{cases}$$

- Negative y's could be allowed.
- The transformation was reported to be successful in transform unimodal skewed distribution into normal distribution, but is not quite useful for bimodal or U-shaped distribution.

► John and Draper(1980) proposed the following modification which they called "Modulus Transformation".

$$y(\lambda) = \begin{cases} \text{Sign}(y)\frac{(|y|+1)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \text{Sign}(y)\log(|y|+1), & \text{if } \lambda = 0, \end{cases}$$

where

$$\text{Sign}(y) = \begin{cases} 1, & \text{if } y \geq 0; \\ -1, & \text{if } y < 0. \end{cases}$$

- Negative y's could be allowed.
- It works best at those distribution that is somewhat symmetric. A power transformation on a symmetric distribution is likely going to introduce some degree of skewness.

▶ Bickel and Doksum(1981) gave the following slight modification in their examination of the asymptotic performance of the parameters in the Box-Cox transformations model:

$$y(\lambda) = \frac{|y|^{\lambda}\mathrm{Sign}(y) - 1}{\lambda}, \qquad \text{for } \lambda > 0,$$

where

$$\mathrm{Sign}(y) = \begin{cases} 1, & \text{if } y \geq 0; \\ -1, & \text{if } y < 0. \end{cases}$$

► Yeo and Johnson(2000) made a case for the following transformation:

$$y(\lambda) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0,\, y \geq 0; \\ \log(y+1), & \text{if } \lambda = 0,\, y \geq 0; \\ \frac{(1-y)^{2-\lambda}-1}{\lambda-2}, & \text{if } \lambda \neq 2,\, y < 0; \\ -\log(1-y), & \text{if } \lambda = 2,\, y < 0. \end{cases}$$

When estimating the transformation parameter, they found the value of λ that minimizes the Kullback-Leibler distance between the normal distribution and the transformed distribution.

The inference on the transformation parameter

▶ The main objective in the analysis of Box-Cox transformation model is to make inference on the transformation parameter λ, and Box and Cox(1964) considered two approaches.

▶ The first approach is to use the Maximum Likelihood method. This method is commonly used since it's conceptually easy and the profile likelihood function is easy to compute in this case. Also it's easy to obtain an approximate CI for λ because of the asymptotic property of MLE.

▶ We assume that transformed responses $\boldsymbol{y}(\lambda) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. We observe the design matrix $\mathbf{X}$ and the raw data $\boldsymbol{y}$, and the model parameters are $(\lambda, \boldsymbol{\beta}, \sigma^2)$.

▶ The density for the $\boldsymbol{y}(\lambda)$ is

$$f(\boldsymbol{y}(\lambda)) = \frac{\exp(-\frac{1}{2\sigma^2}(\boldsymbol{y}(\lambda) - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y}(\lambda) - \mathbf{X}\boldsymbol{\beta}))}{(2\pi\sigma^2)^{\frac{n}{2}}}.$$

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the transformation from $\boldsymbol{y}$ to $\boldsymbol{y}(\lambda)$, then the density for $\boldsymbol{y}$ (which is also the likelihood for the whole model) is

$$L(\lambda, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \mathbf{X}) = f(\boldsymbol{y}) = \frac{\exp(-\frac{1}{2\sigma^2}(\boldsymbol{y}(\lambda) - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y}(\lambda) - \mathbf{X}\boldsymbol{\beta}))}{(2\pi\sigma^2)^{\frac{n}{2}}} J(\lambda, \boldsymbol{y}).$$

▶ To obtain the MLE from the last likelihood equation, we observe that for each fixed λ, the likelihood equation is proportional to the likelihood equation for estimating $(\boldsymbol{\beta}, \sigma^2)$ for observed $\boldsymbol{y}(\lambda)$. Thus the MLE's for $(\boldsymbol{\beta}, \sigma^2)$ are

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(\lambda) &= (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}\boldsymbol{y}(\lambda), \\ \hat{\sigma^2}(\lambda) &= \frac{\boldsymbol{y}(\lambda)'(\mathbf{I}_n - \mathbf{G})\boldsymbol{y}(\lambda)}{n}, \end{aligned}$$

where $\mathbf{G} = \text{ppo}(\mathbf{X}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$.

▶ Substitute $\tilde{\boldsymbol{\beta}}(\lambda)$ and $\hat{\sigma^2}(\lambda)$ into the likelihood equation, and note that for the original form of the Box-Cox transformation, $J(\lambda, \boldsymbol{y}) = \prod_{i=1}^{n} y_i^{\lambda-1}$, we could obtain the profile log likelihood(i.e., the likelihood function maximized over $(\boldsymbol{\beta}, \sigma^2)$) for λ alone.

- $$\begin{aligned} l_P(\lambda) &= l(\lambda|\boldsymbol{y}, \mathbf{X}, \tilde{\boldsymbol{\beta}}(\lambda), \hat{\sigma^2}(\lambda)) \\ &= C - \frac{n}{2}\log(\hat{\sigma^2}(\lambda)) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i) \end{aligned}$$

- Let g be the geometric mean of the response vector(i.e., $g = (\prod_{i=1}^{n} y_i)^{\frac{1}{n}}$), also let $\boldsymbol{y}(\lambda, g) = \frac{\boldsymbol{y}(\lambda)}{g^{\lambda-1}}$. Then it's easy to see

$$l_P(\lambda) = C - \frac{n}{2}\log(s_\lambda^2),$$

where s_λ^2 is the residual sum of squares divided by n from fitting the linear model $\boldsymbol{y}(\lambda, g) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. So to maximize the profile log-likelihood, we only need to find a λ that minimizes $s_\lambda^2 = \frac{\boldsymbol{y}(\lambda,g)'(\mathbf{I}_n - \mathbf{G})\boldsymbol{y}(\lambda,g)}{n}$.

▶ Without any further effort and just use the standard likelihood methods, we could easily give a likelihood ratio test. For test $H_0 : \lambda = \lambda_0$, the test statistic is $W = 2[l_P(\hat{\lambda}) - l_P(\lambda_0)]$. Asymptotically W is distributed as χ^2_1. Carefully note that W is a function of both the data (through $\hat{\lambda}$) and λ_0.

▶ A large sample CI for λ is easily obtainable by inverting the likelihood ratio test. Let $\hat{\lambda}$ be the MLE of λ, then an approximate $(1-\alpha)100\%$ CI for λ is

$$\{\lambda \mid n \times \log(\frac{\text{SSE}(\lambda)}{\text{SSE}(\hat{\lambda})}) \leq \chi^2_1(1-\alpha)\},$$

where $\text{SSE}(\lambda) = \boldsymbol{y}(\lambda, g)'(\mathbf{I}_n - \mathbf{G})\boldsymbol{y}(\lambda, g)$. The accuracy of the approximation is given by the following fact:

$$P(W \leq \chi^2_1(1-\alpha)) = 1 - \alpha + O(n^{-\frac{1}{2}}).$$

► It's also not hard to derive a test using Rao's score statistic. Atkinson(1973) first proposed a score-type statistic for test $H_0 : \lambda = \lambda_0$, although the derivation were not based on likelihood theory. Lawrence(1987) modified the result by Atkinson(1973), by employing the standard likelihood theory.

▶ The second approach outlined in Box and Cox(1964) is to use Bayesian method. In this approach, we need to first ensure that the model is fully identifiable. If $\mathbf{X}$ is not of full column rank, then $\boldsymbol{\beta}$ is not estimable(or more accurately identifiable). So we further assume $\mathbf{X}$ is $n \times p$ matrix and $\mathrm{rank}(\mathbf{X}) = r(r \leq p)$. Now using the full-rank factorization to write $\mathbf{X} = \mathbf{AR}$(Nalini and Day, p40, result 2.2.1), it's easy to reparameterize the model as $\boldsymbol{y}(\lambda) \sim \mathrm{N}(\mathbf{A}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n)$, where $\mathbf{A} : n \times r$ is of full column rank and $\boldsymbol{\theta} = \mathbf{R}\boldsymbol{\beta}$ is itself estimable.

▶ We now consider the prior distribution for the parameters $(\lambda, \boldsymbol{\theta}, \sigma^2)$. Box and Cox(1964) propose the following prior

$$\pi_1(\lambda, \boldsymbol{\theta}, \sigma) \propto \pi(\lambda) \times \frac{1}{\sigma} \times \frac{1}{g^{(\lambda-1)r}},$$

where g is the geometric mean of response vector $\boldsymbol{y}$, and $\pi(\lambda)$ is some prior distribution for λ only.

▶ Pericchi(1981) considered another joint prior distribution

$$\pi_2(\lambda, \boldsymbol{\theta}, \sigma) \propto \pi(\lambda) \times \frac{1}{\sigma^{r+1}},$$

again $\pi(\lambda)$ is some prior distribution for λ only.

▶ So what's the rationale of choosing such prior distributions?

▶ When $\lambda = 1$(i.e., no transformation performed), the $\boldsymbol{\theta}$ is a location parameter and σ is a scale parameter, so the natural non-informative prior for $\boldsymbol{\theta}$ and σ should be uniform and $\frac{1}{\sigma}$ respectively. This implies

$$\pi_1(\lambda = 1, \boldsymbol{\theta}, \sigma) = p(\boldsymbol{\theta}, \sigma | \lambda = 1) \times \pi(\lambda = 1) \propto \pi(\lambda = 1) \times \frac{1}{\sigma}.$$

Box and Cox(1964) then assumes that the transformation is approximately linear over the range of observations, that is

$$\mathrm{E}(y_i(\lambda)) \approx a_\lambda + b_\lambda \mathrm{E}(y_i),$$

where b_λ is some representative of the gradient $\frac{dy(\lambda)}{dy}$. This implies that when $\lambda \neq 1$, each element of $\boldsymbol{\theta}$ is multiplies by a scale of b_λ. So the prior for $\boldsymbol{\theta}$ when $\lambda \neq 1$ should be $\frac{1}{|b_\lambda|^r}$.

▶ Box and Cox(1964) chose $b_\lambda = J(\lambda, \boldsymbol{y})^{\frac{1}{n}} = g^{\lambda-1}$, which they admitted that such choice was "somewhat arbitrary". This gives the Box-Cox version of the prior distribution.

▶ Pericchi(1981) followed exactly the same argument, with the exception that the use of Jefferys' prior for $(\boldsymbol{\theta}, \sigma)$ instead of invariant non-informative prior.

▶ Clearly the Box and Cox's prior is "outcome-dependent", which seems to be an undesirable property.

▶ It's not hard to see that the posterior distribution for Box and Cox prior is

$$\pi_1(\lambda, \boldsymbol{\theta}, \sigma | \boldsymbol{y}, \mathbf{A}) \propto \frac{1}{\sigma^{n+1}} \times \exp(-\frac{S_\lambda + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{A}' \mathbf{A} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})}{2\sigma^2}) \times g^{(\lambda-1)(n-r)} \times \pi(\lambda),$$

where $S_\lambda = (\boldsymbol{y}(\lambda) - \mathbf{A}\hat{\boldsymbol{\theta}})'(\boldsymbol{y}(\lambda) - \mathbf{A}\hat{\boldsymbol{\theta}})$.

The posterior distribution for Pericchi's prior is

$$\pi_2(\lambda, \boldsymbol{\theta}, \sigma | \boldsymbol{y}, \mathbf{A}) \propto \frac{1}{\sigma^{n+r+1}} \times \exp(-\frac{S_\lambda + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})' \mathbf{A}' \mathbf{A} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})}{2\sigma^2}) \times g^{(\lambda-1)n} \times \pi(\lambda).$$

► Integrating out $(\boldsymbol{\theta}, \sigma)$, we can then get the posterior log likelihood for λ alone. For Box and Cox's prior,

$$l_1(\lambda) = C - \frac{1}{2}(n-r)\log(\frac{S_\lambda}{n-r} \times \frac{1}{g^{2(\lambda-1)}}),$$

and for Pericchi's prior,

$$l_2(\lambda) = C - \frac{1}{2}n\log(\frac{S_\lambda}{n} \times \frac{1}{g^{2(\lambda-1)}}).$$

► One may note that the posterior log likelihood based on Pericchi's prior is the same as the profile log likelihood from the maximum likelihood method. So, they will lead to identical inference about λ.

► Using the normal approximation to posterior density, we can derive the approximate $100(1-\alpha)\%$ HPD credible set as follows:

$$\begin{aligned}
\mathrm{HPD}_1 &= \{\lambda \mid (n-r)\times\log(\frac{\mathrm{SSE}(\lambda)}{\mathrm{SSE}(\hat{\lambda})}) \le \chi_1^2(1-\alpha)\}, \\
\mathrm{HPD}_2 &= \{\lambda \mid n\times\log(\frac{\mathrm{SSE}(\lambda)}{\mathrm{SSE}(\hat{\lambda})}) \le \chi_1^2(1-\alpha)\},
\end{aligned}$$

where $\mathrm{SSE}(\lambda) = \boldsymbol{y}(\lambda, g)'(\mathbf{I}_n - \mathbf{G})\boldsymbol{y}(\lambda, g)$.

- Note the HPD_2 is the same as the confidence interval we derived using the maximum likelihood mathod.
- $\mathrm{HPD}_2 \subseteq \mathrm{HPD}_1$.

► A transformation parameter could also be estimated on the basis of enforcing a particular assumption of the linear model.

► For example, if we want to ensure the additivity(or linearity) in the linear model, we could select a transformation parameter that will minimize the F-value for the degree of freedom for non-additivity. This idea was firstly expressed in Tukey(1949).

► As a particular example, consider the example in Professor Chen's class notes for Stat321, note 8, p41. Suppose that we face two competing models $\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, and $\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$, the latter clearly being non-linear. An estimate of λ, based the above argument, should be the one that minimize the F-statistic that associated with the following model comparison test(which is a likelihood ratio test):

$$\begin{cases} H_0 : & \mathrm{E}(y(\lambda)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\ H_1 : & \mathrm{E}(y(\lambda)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2, \\ & \text{not all } \ \beta_3, \beta_4, \beta_5 \ \text{ are } \ 0. \end{cases}$$

► Professor Chen's illustration is based on the scaled Box-Cox transformation, but it's equivalent to using our version of transformation.

Some cautionary notes on using the Box-Cox transformations

▶ The usual practice when using Box-Cox transformation is to first estimate λ. Then the estimated λ is viewed as *known*, and analysis (points estimates, confidence intervals) are carried out in the usual fashion.

▶ In the original paper of Box and Cox(1964), their suggestion was to "fix one, or possibly a small number, of λs and go ahead with the detailed estimation". In their examples, they used what's usually called "snap to the grid" methods to choose the estimate of λ.

▶ In this approach, we are essentially making inference about $(\boldsymbol{\beta}, \sigma)$ conditioning on $\lambda = \hat{\lambda}$.

▶ Bickel and Doksum(1981) studied the joint estimation of $(\lambda, \boldsymbol{\beta})$. They proved, and illustrated through numerical example, that the asympototic marginal(unconditional) variance of $\hat{\boldsymbol{\beta}}$ could be inflated by a very large factor over the conditional variance for fixed λ.

▶ Much research have been done after Bickel and Doksum(1981), either on a philosophical or on a technical level. Although there does not appear to be any definite result, most research agree that while there is an effect on not knowing the true value of λ, it's cost may not be large enough to discredit the conventional application based on conditioning.

▶ One partial remedy to the problem is to use the scaled form of Box-Cox transformation:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}\frac{1}{g^{\lambda-1}}, & \text{if } \lambda \neq 0; \\ \log y \times g, & \text{if } \lambda = 0, \end{cases}$$

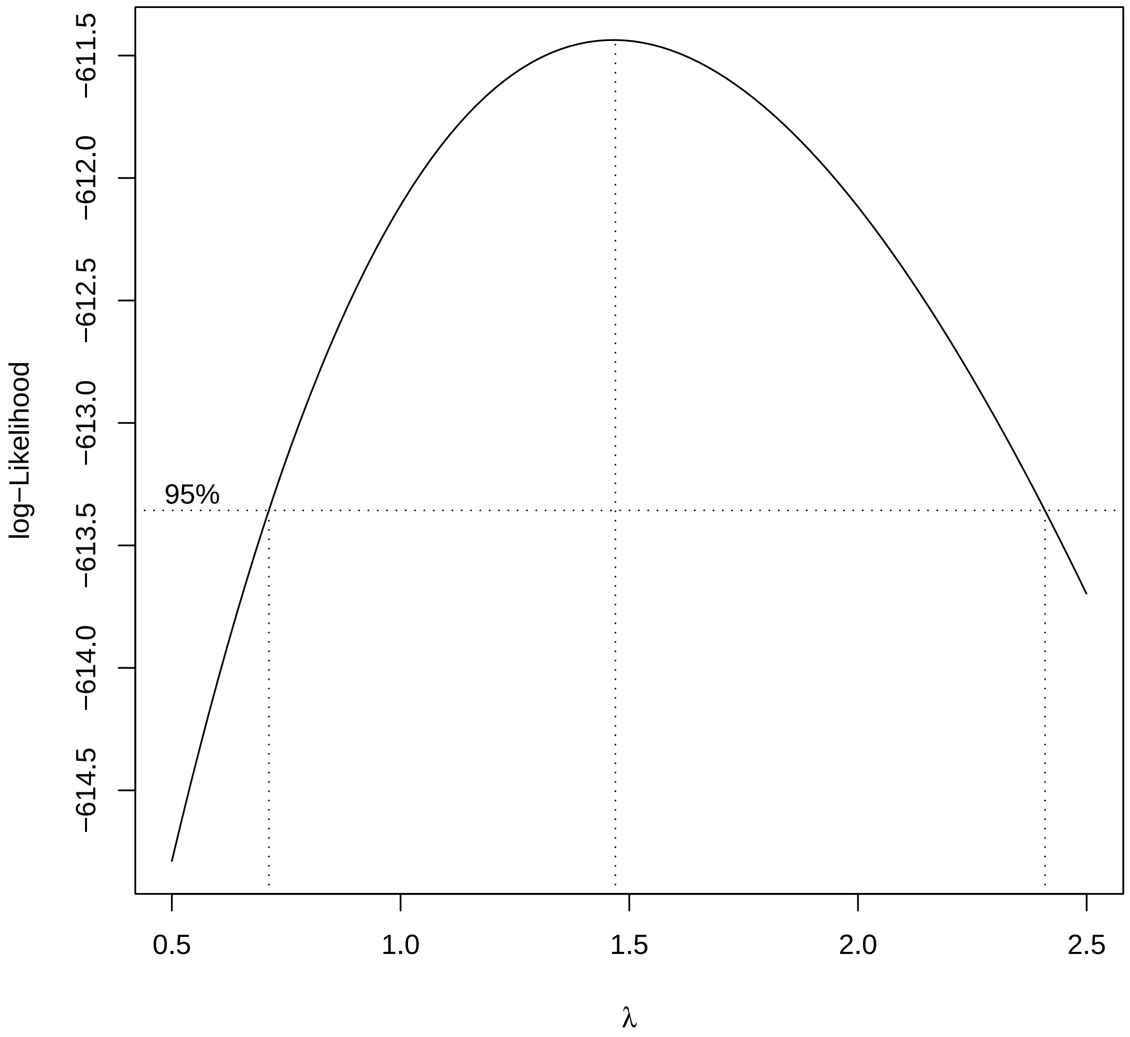
where g, as before, is the geometric mean of the response vector.

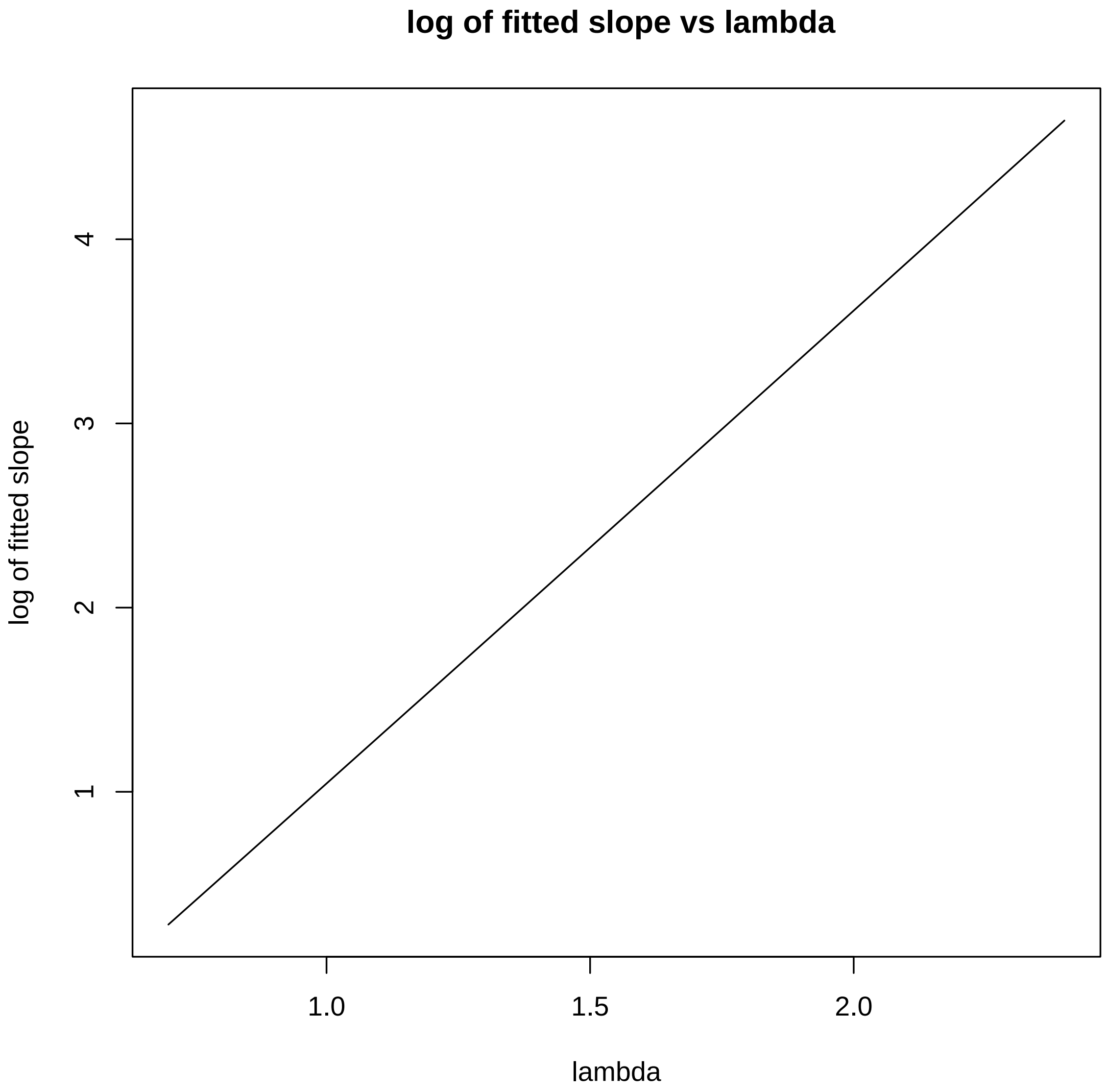
- Such scaling can effectively control the size of transformed responses, and can also reduce the conditional variance for $\boldsymbol{\beta}$.
- The t,F statistic on the scaled responses will be the same as those from unscaled responses.

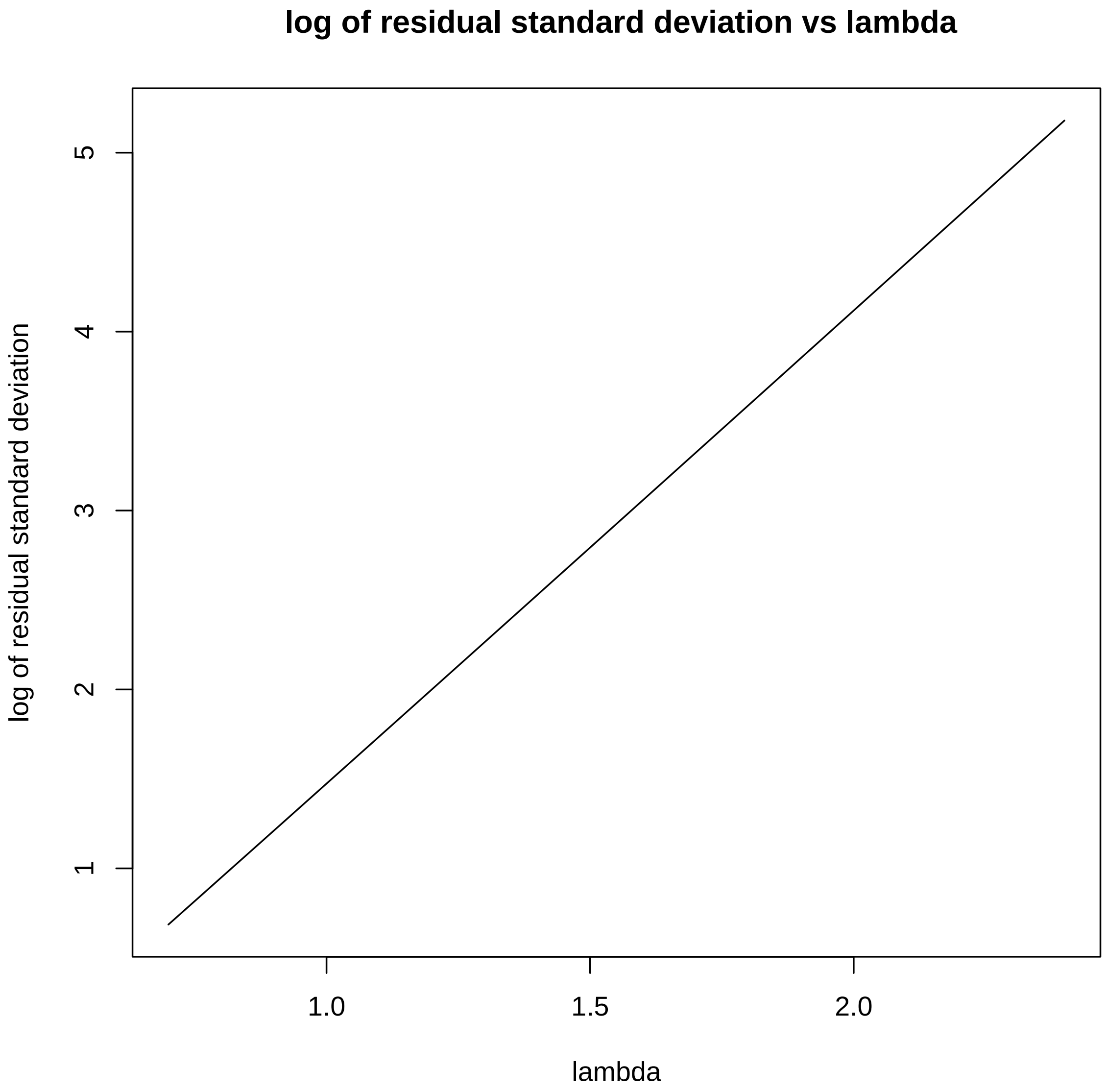
▶ In the most recent paper by Chen, Lockhart and Stephens(2002), they claim that the ratio parameter $\phi = \frac{\boldsymbol{\beta}}{\sigma}$ is the one that is "physically meaningful" in the Box-Cox transformation model. They showed, in particular, that the MLE of the ratio of slope to residual standard deviation is consistent and relatively stable.

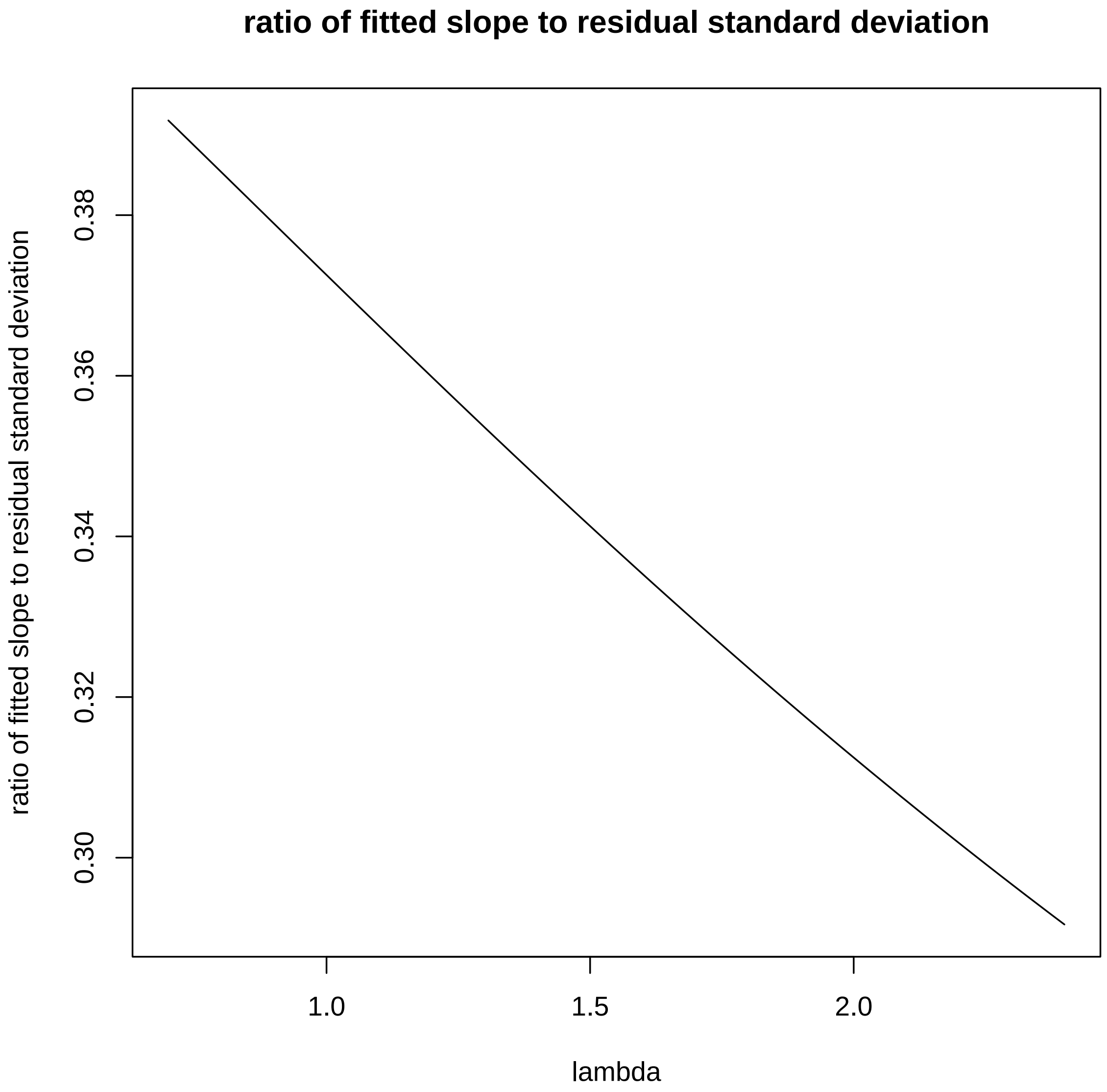
▶ We now consider a simple data set in Chen, Lockhart and Stephens(2002) that contains 107 pairs of observations (y, x). y measures the distance driven(in Km) and x measures the amounts of gas consumed(in litres). They applied the Box-Cox transformation to y's and fitted a simple linear regression of transformed y on x.

▶ We first obtain a profile log-likelihood plot. It's clear that a sensible estimate for λ should be 1.5, and a 95% CI for λ could be from approximately 0.7 to 2.4.


log−Likelihood
−611.5
−612.0
−612.5
−613.0
−613.5
−614.0
−614.5
95%
0.5
1.0
1.5
2.0
2.5
λ


log of fitted slope vs lambda
log of fitted slope
1
2
3
4
1.0
1.5
2.0
lambda


log of residual standard deviation vs lambda
log of residual standard deviation
1
2
3
4
5
1.0
1.5
2.0
lambda


ratio of fitted slope to residual standard deviation
ratio of fitted slope to residual standard deviation
0.30
0.32
0.34
0.36
0.38
1.0
1.5
2.0
lambda

- Estimates for selected λ.

λ	0.7	1.0	1.46	1.5	2.0
$\hat{\beta}$	1.90	11.09	167.30	211.93	4097.98
$\hat{\sigma}$	4.85	29.78	486.75	620.98	13114.89
$\frac{\hat{\beta}}{\hat{\sigma}}$	0.392	0.373	0.344	0.341	0.312

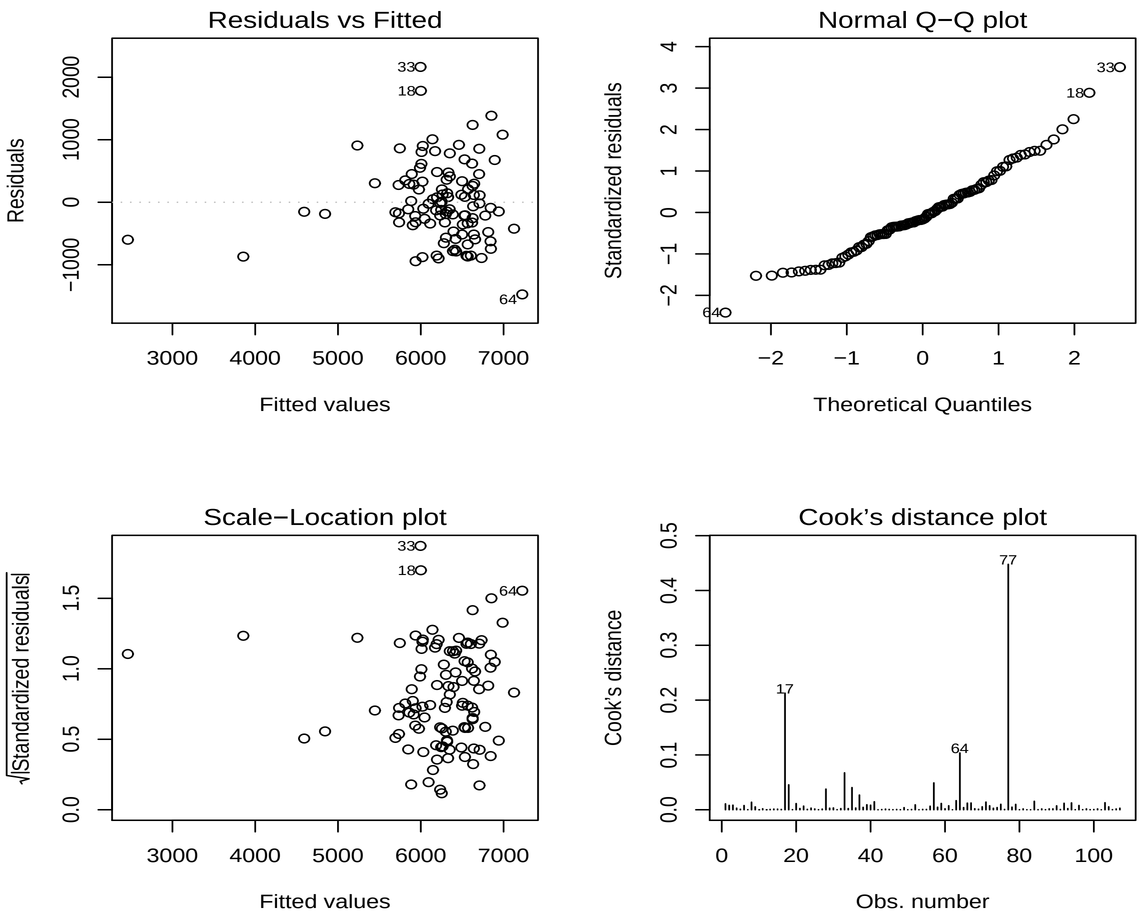
Table 1: Estimates for selected λ.

- Fitted model information for $\lambda = 1.5$.

```
> summary(fit.boxcox)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -636.87     686.25  -0.928    0.356
liter         211.93      21.07  10.058   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 621 on 105 degrees of freedom
Multiple R-Squared: 0.4907,     Adjusted R-squared: 0.4858
F-statistic: 101.2 on 1 and 105 DF,  p-value: < 2.2e-16
```

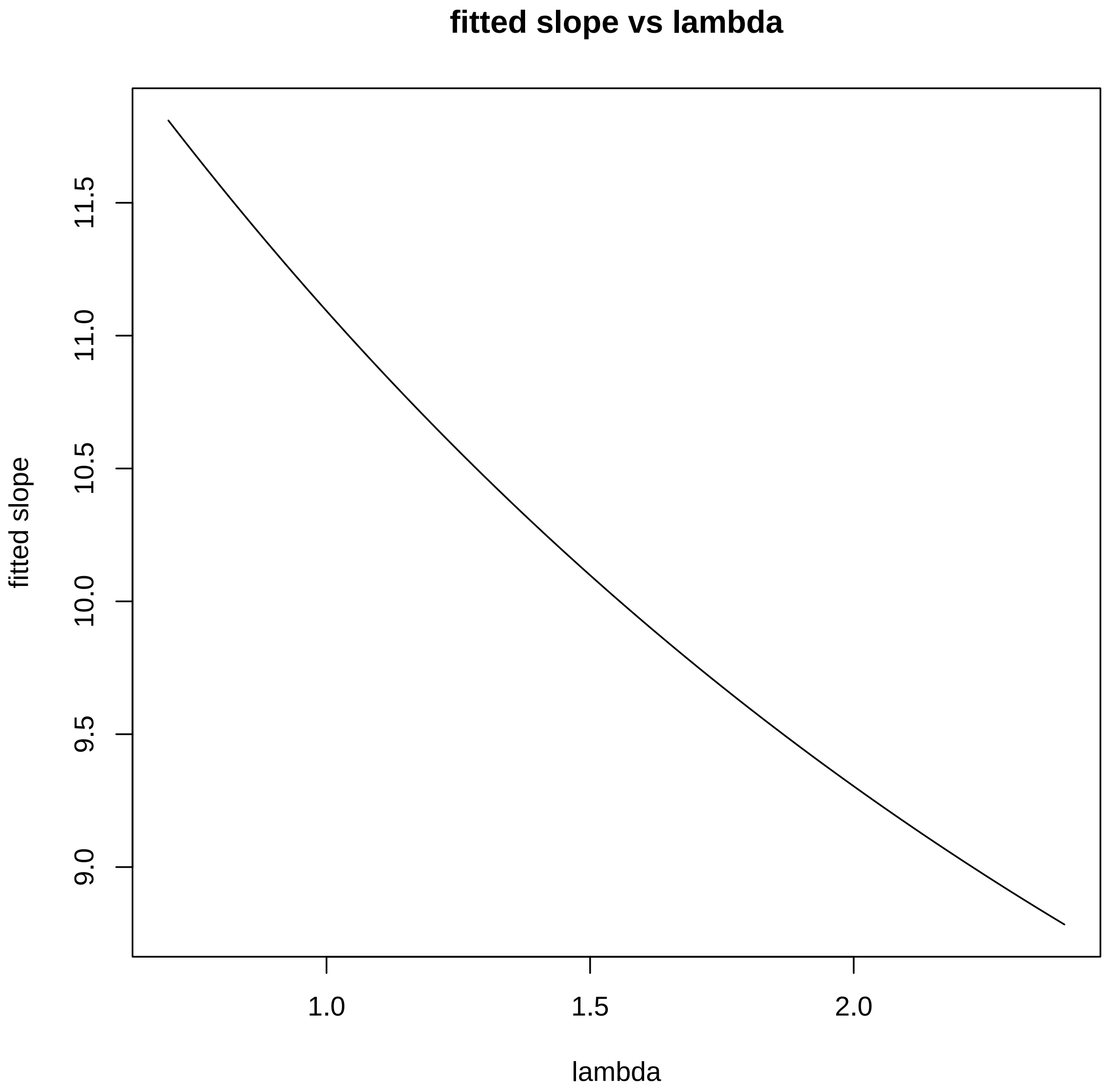
- Model diagnostic plots.

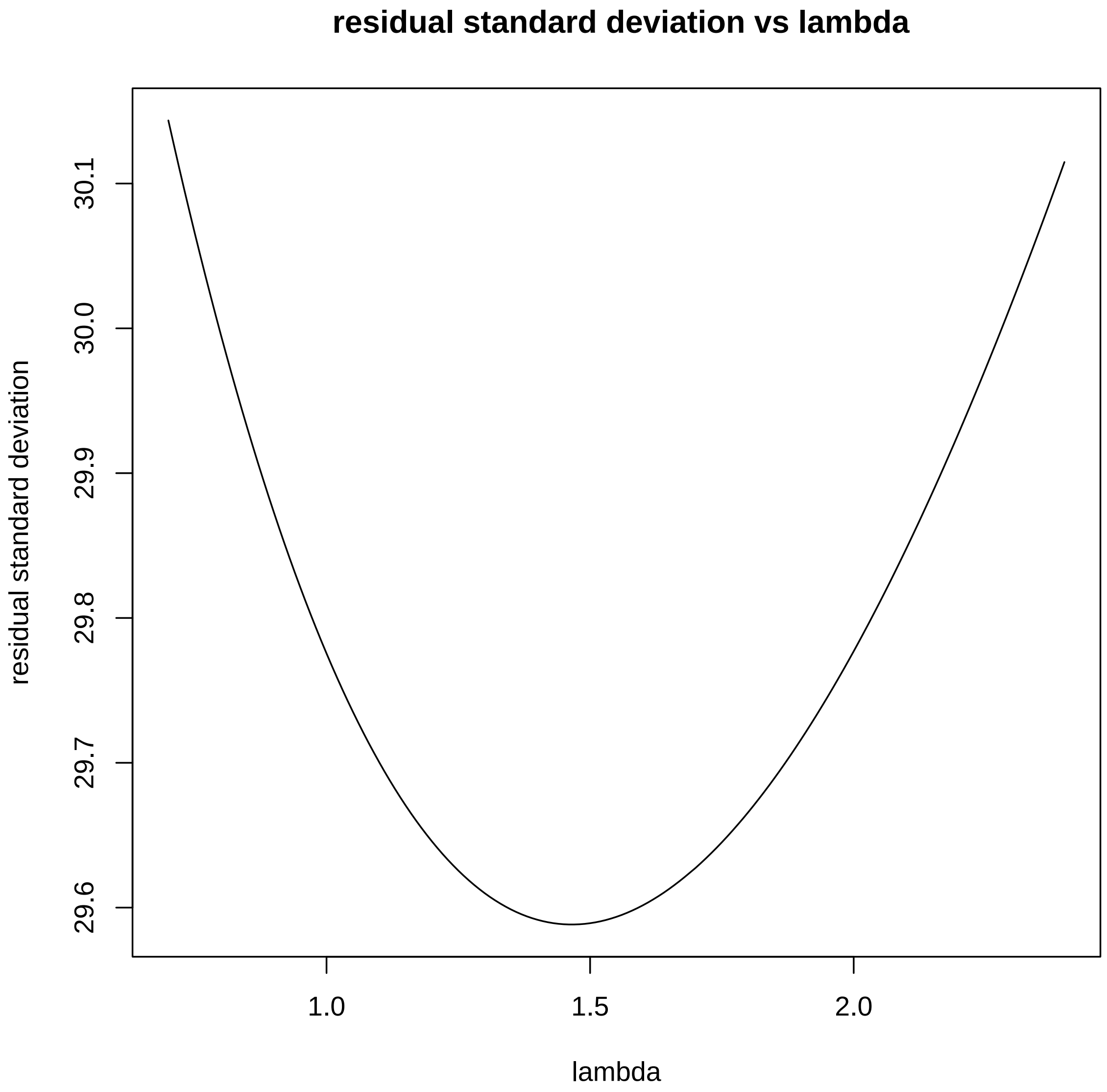


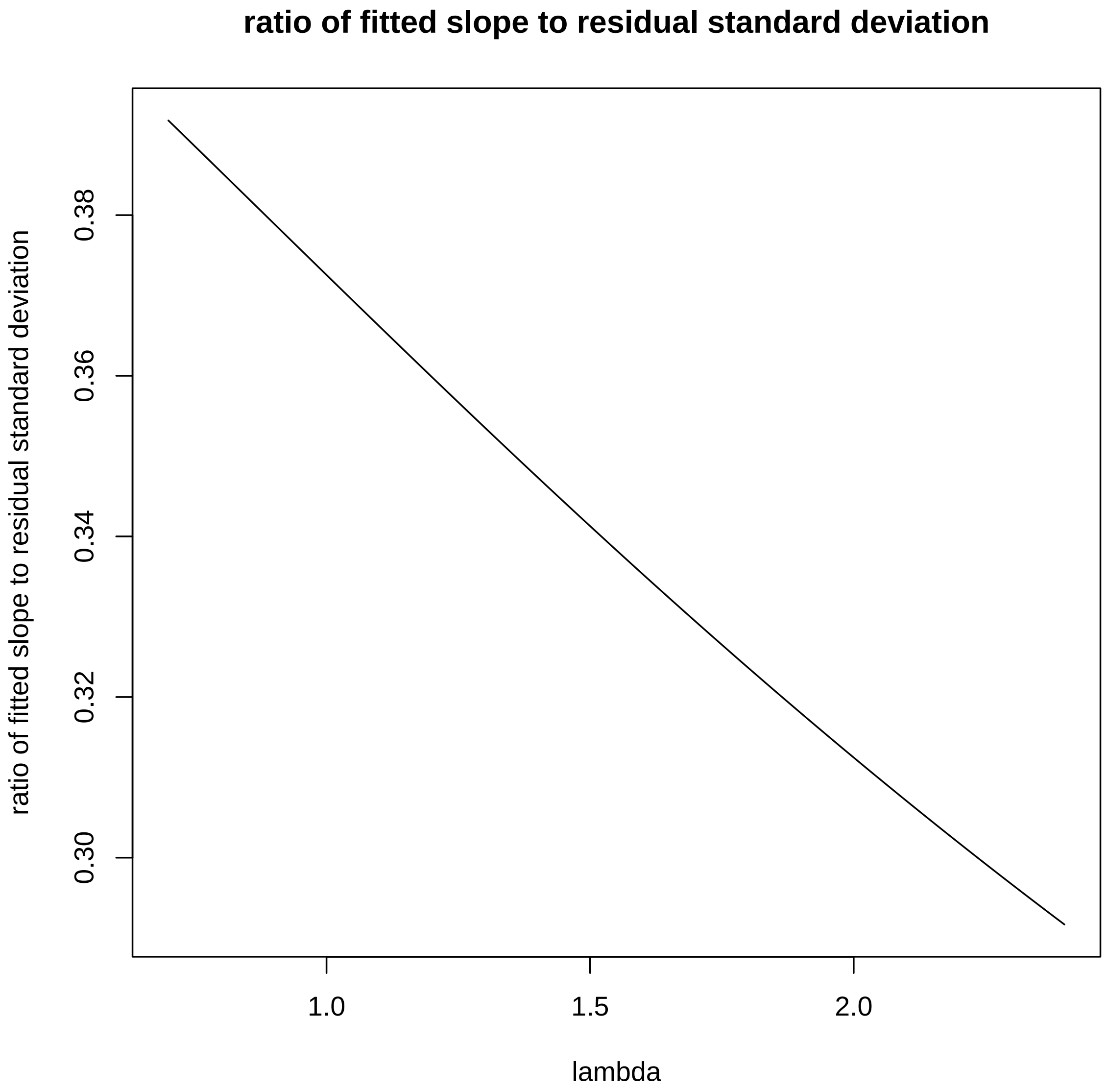
► We now repeated the same analysis for the scaled Box-Cox transformation.

λ	0.7	1.0	1.46	1.5	2.0
$\hat{\beta}$	11.81	11.09	10.17	10.10	9.30
$\hat{\sigma}$	30.14	29.78	29.59	29.59	29.78
$\frac{\hat{\beta}}{\hat{\sigma}}$	0.392	0.373	0.344	0.341	0.312

Table 2: Estimates for selected λ for scaled transformation.


fitted slope vs lambda
fitted slope
9.0
9.5
10.0
10.5
11.0
11.5
1.0
1.5
2.0
lambda


residual standard deviation vs lambda
residual standard deviation
29.6
29.7
29.8
29.9
30.0
30.1
1.0
1.5
2.0
lambda

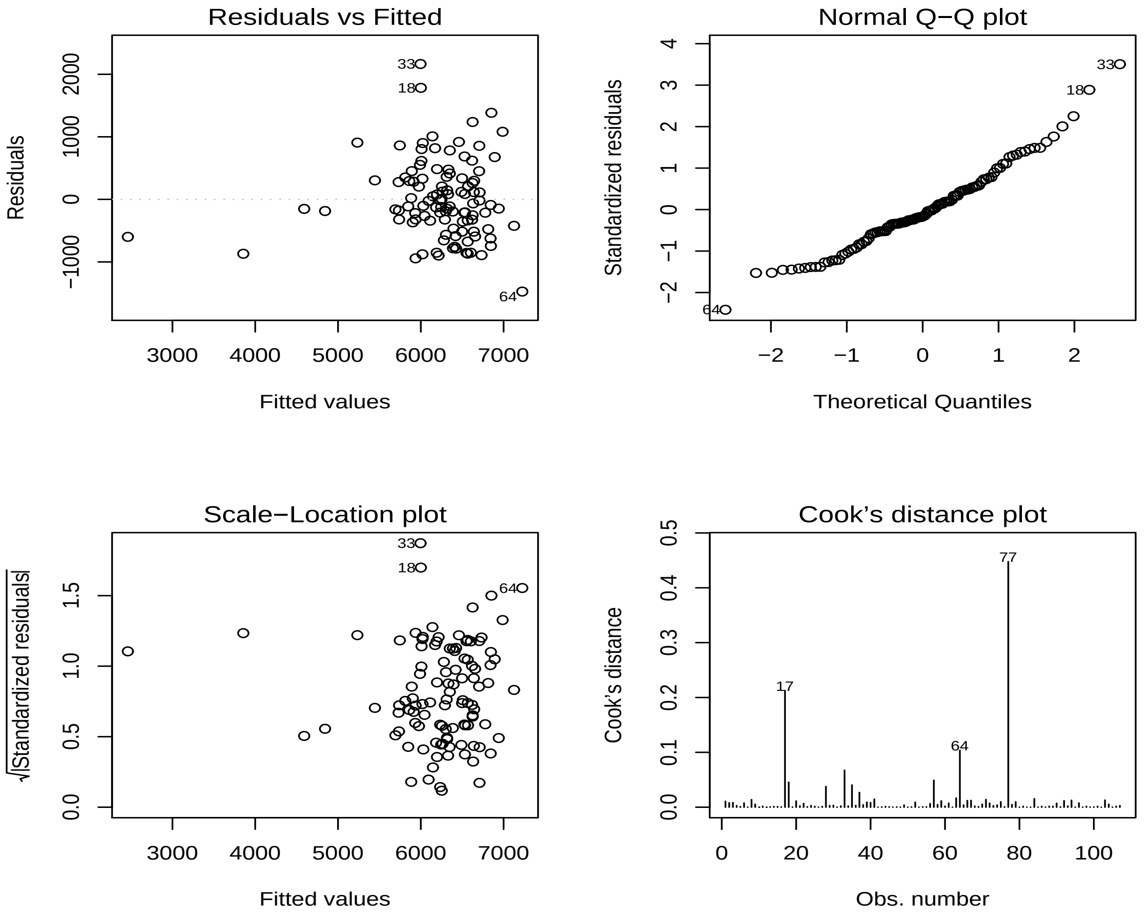

ratio of fitted slope to residual standard deviation
ratio of fitted slope to residual standard deviation
0.30
0.32
0.34
0.36
0.38
1.0
1.5
2.0
lambda

▶ Fitted model information for $\lambda = 1.5$, scaled.

```
> summary(fit.boxcox.scale)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -30.346     32.699  -0.928    0.356
liter         10.098      1.004  10.058   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.59 on 105 degrees of freedom
Multiple R-Squared: 0.4907,     Adjusted R-squared: 0.4858
F-statistic: 101.2 on 1 and 105 DF,  p-value: < 2.2e-16
```

- Model diagnostic plots.



▶ So what can we learn from this example?

▶ For unscaled transformation, the estimates seemingly vary greatly with the very minor change of choice of λ. But this is not as bad as it may look like!

▶ By choosing $\lambda = 1.46$ and $\lambda = 1.5$, we end up with two fitted equations:

$$\begin{aligned} \hat{y}(\mathrm{Km}^{1.46}) &= -406.35(\mathrm{Km}^{1.46}) + 167.30(\frac{\mathrm{Km}^{1.46}}{\mathrm{liter}})x(\mathrm{liter}), \\ \hat{y}(\mathrm{Km}^{1.50}) &= -636.87(\mathrm{Km}^{1.50}) + 211.93(\frac{\mathrm{Km}^{1.50}}{\mathrm{liter}})x(\mathrm{liter}). \end{aligned}$$

Thus you can compare 167.30 to 211.93 but not $167.30(\frac{\mathrm{Km}^{1.46}}{\mathrm{liter}})$ to $211.93(\frac{\mathrm{Km}^{1.50}}{\mathrm{liter}})$.

▶ This idea is one of the major arguments of Box and Cox(1982) and Hinkley and Runger(1984) in their rebuttal of Bickel and Doksum(1981). They call this consideration "scientific relevance".

▶ If we consider the proposal by Chen, Lockhart and Stephens(2002) using the ratio parameter $\phi = \frac{\boldsymbol{\beta}}{\sigma}$, we can see that the physical scale of ϕ is invariant under transformation.

▶ If we are using scaled Box-Cox transformation, the fitted model (for $\lambda = 1.46$ and $\lambda = 1.5$) would be

$$\begin{aligned} \hat{y}(\text{Km}) &= -24.701(\text{Km}) + 10.170(\frac{\text{Km}}{\text{liter}})x(\text{liter}), \\ \hat{y}(\text{Km}) &= -30.346(\text{Km}) + 10.098(\frac{\text{Km}}{\text{liter}})x(\text{liter}). \end{aligned}$$

Thus comparing the coefficients now makes practical sense.

- ► However, such scaling has not much effect other than producing comparable estimates. The test statistic remain the same as if unscaled. Thus it may not be useful for ANOVA model.

What's beyond this overview?

- Other methods of estimation: e.g., robust estimator, non-parametric estimator.
- Estimation procedure on more structured data: e.g., mixed models, missing data.
- Continued examination of what's the most appropriate way to conduct post transformation analysis.